文章编号:1001 - 9081(2008)03 - 0792 - 03

基于 GMM 的普通话和四川方言独立文本的说话人确认

赵 靖,龚卫国,杨利平
(重庆大学 光电技术及系统教育部重点实验室,重庆 400030)
(wggong@ cqu. edu. cn)

摘 要:针对训练和测试阶段中的语音数据类型(普通话和四川方言)的不匹配导致说话人确认系统性能下降很大的问题,提出了一种新的建立高斯混合模型(GMM)方法——普通话和四川方言按比例混合建立普通话和四川方言联合 GMM 的方法,并发现使系统针对普通话和四川方言不匹配导致的性能下降率至很低(2.79%)的比例。实验结果表明,该方法可以有效地加强测试阶段针对语种变化的鲁棒性,可以有效的减少普通话和四川方言在训练和测试阶段的不匹配造成的性能下降率。

关键词:说话人确认;高斯混合模型;独立文本;双语种说话人确认
中图分类号: TP391.4 **文献标志码**:A

Mandarin-Sichuan dialect bilingual text-independent speaker verification using GMM

ZHAO Jing, GONG Wei-guo, YANG Li-ping
(*Key Laboratory for Optoelectronic Technology and System of Education Ministry of China, Chongqing University, Chongqing* 400030, *China*)

Abstract: Due to the mismatch between mandarin and Sichuan dialect in training and test stages, the performance of speaker verification system degrades dramatically. To solve this problem, a combined Gaussian Mixture Model (GMM), which is trained by proportional pooling mandarin and Sichuan dialect, was presented in this paper. Compared with the Gaussian mixture model trained solely using mandarin/Sichuan dialect, the combined Gaussian mixture model described the characteristic of speaker from both mandarin and Sichuan dialect. Experiments on a self-built mandarin-Sichuan dialect speech database demonstrate that the introduced combined Gaussian mixture model is more robust for speech mismatching between mandarin and Sichuan dialect. A proper proportion between pooling mandarin and Sichuan dialect speech was also provided.

Key words: speaker verification; Gaussian Mixture Models (GMM); text-independent; bilingual speaker verification

0 引言

说话人确认是根据说话人的输入语音鉴别原说话人身份的过程[1]。随着计算机技术的不断进步,通过说话人的语音进行身份识别将使识别过程更加方便、有效和安全。

语音识别领域的语音库大部分都是基于单语种的,如 SIVA 语音库,NTT 语音库,TIMIT 语音库,NTIMIT 语音库[2]574,而用来做中国方言研究的语音数据库却很少。文献[3]通过研究发现,测试阶段的语种和训练阶段的语种的不匹配将导致很大的系统性能下降率,当训练语音用中国普通话和越南话而测试语音用美国英语时下降率更明显。

中国以普通话为全国的通用语言,而中国是一个有很多方言的国家,其中四川方言在西南地区有很大的影响。四川方言声母比普通话少,还有韵母儿化现象,只有舌尖前音,没有舌尖后音,普通话与四川方言在发音方式上有很大的区别。

在本文的研究中,提出了一种新的建立高斯混合模型(Gaussian Mixture Models,GMM)方法——普通话和四川方言按比例混合建立普通话和四川方言联合 GMM 的方法,研究发现了使系统针对普通话和四川方言不匹配导致的性能下降率至很低(2.79%)的比例,从而开发了普通话—四川方言独立文本说话人确认系统。实验结果表明,用这种普通话和四川方言按照比例混合建立的联合 GMM 是一种比用传统单一语种训练的 GMM 更健壮的模型,并且可以很好的克服训练语音和测试语音(普通话和四川方言)的不匹配带来的性能下降问题。

1 高斯混合模型(GMM)

高斯混合模型是用多个高斯分布的概率密度函数的组合来描述特征矢量在概率空间的分布状况。每个说话人对应一个 GMM[2]572 - 573。

该模型用多个具有高斯分布的概率密度的加权和来表示,该概率密度函数的个数称之为高斯模型的混合数。一个具有 M 个混合数的 d 维 GMM,可以表示为:

$$P(x \mid \lambda) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \tag{1}$$

其中:$\boldsymbol{x}$ 为 d 维观察矢量;ω_i 为混合权重,且满足 $\sum_{i=1}^{M} \omega_i = 1$;$p_i(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ 为 d 维高斯函数,表示 GMM 模型的第 i 个高斯分量,$\boldsymbol{\mu}_i$ 为该高斯分量的均值矢量,$\boldsymbol{\Sigma}_i$ 为协方差矩阵。整个 GMM 由各混合分量的均值矢量、协方差矩阵以及混合权重来描述,用 λ 来表示该模型,有:

$$\lambda = \{\omega_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\};\ i = 1,2,\cdots,M \tag{2}$$

GMM 模型参数的训练一般采用最大似然估计(Maximum Likelihood Estimation, MLE)的方法。假设某说话人训练的观

收稿日期:2007 - 10 - 09;**修回日期**:2007 - 12 - 07。

作者简介:赵靖(1981 -),男,山东潍坊人,硕士研究生,主要研究方向:信息获取与处理、人工智能; 龚卫国(1957 -),男,重庆人,教授,博士生导师,主要研究方向:模式识别及机器视觉、智能化信息技术及系统。

察矢量序列 $\boldsymbol{X}=\{\boldsymbol{x}_t \mid t=1,2,\cdots,T\}$ 中各观察矢量 $\boldsymbol{x}_t$ 是独立不相关的,对于 GMM 模型 λ 的似然度可表示为:

$$L(\lambda \mid \boldsymbol{x}) = p(\boldsymbol{x} \mid \lambda) = \prod_{t=1}^{T} P(x_i \mid \lambda) \tag{3}$$

训练的目的是找到一组模型参数 $\hat{\lambda}$,使得 $L(\lambda \mid X)$ 最大,即:

$$\hat{\lambda} = \arg\max_{\lambda} L(\lambda \mid X) = \arg\max_{\lambda} P(X \mid \lambda) \tag{4}$$

高斯混合模型已经证明对单语种有很好的识别率,但是将高斯混合模型应用于多种语种的说话人识别时,会遇到测试数据与训练数据不匹配导致说话人识别系统性能下降问题。这说明 GMM 在处理多语种说话人识别方面还存在不足[3]443,而本文提出的按不同比例联合普通话和四川方言的高斯混合模型,可以很好的解决上述提出的问题。

2　按不同比例混合的联合高斯混合模型

传统的 GMM 都是用单一语种来训练和测试,而用两种或者多种语种进行说话人识别时会遇到因为训练数据与测试数据的不同而带来的系统性能下降率问题[5]293。本文将普通话与四川方言分别作为训练数据和测试数据,同样遇到了这个问题。

用普通话训练模型而分别用普通话和四川方言测试,分别得到的错误率[6-7]为 EER_{MT},EER_{DT},文献[5]指出,性能下降率可以用式(5)表示:

$$\text{系统性能下降率} = \frac{EER_{\mathrm{DT}} - EER_{\mathrm{MT}}}{EER_{\mathrm{MT}}} \tag{5}$$

为了解决上述提出的语种依赖性带来的性能下降问题,本文用一种按比例混合普通话和四川方言来建立联合 GMM 的方法来做实验,本文定义比例系数如下:

$$\alpha = \frac{\text{方言句子时间总长度}}{\text{普通话句子时间总长度}} \tag{6}$$

分别令 α 等于 0/10(只用普通话建立模型),2/8,4/6,5/5,6/4,8/2,10/0(只用四川方言建立模型),从每个说话人的 20 条语音(普通话和四川方言各 10 条)中随机选择按照比例 α 混合,并用其来训练按不同比例混合普通话和四川方言的联合 GMM。

普通话和四川方言联合 GMM 参数的训练仍然采用最大似然估计方法,但是说话人训练的观察矢量序列 $\boldsymbol{X}_{\mathrm{Pooled}}=\{x_t;\ t=1,2,\cdots,T\}$ 中观察矢量 $\boldsymbol{X}_{\mathrm{Pooled}}$ 时包含了普通话和四川方言的语音特征,对于普通话和四川方言联合 GMM 模型 $\lambda_{\mathrm{Pooled}}$ 的似然度可表示为:

$$L(\lambda_{\mathrm{Pooled}} \mid \boldsymbol{X}_{\mathrm{Pooled}}) = P(\boldsymbol{X}_{\mathrm{Pooled}} \mid \lambda_{\mathrm{Pooled}}) = \prod_{t=1}^{T} P(x_t \mid \lambda_{\mathrm{Pooled}}) \tag{7}$$

找到一组模型参数 $\hat{\lambda}_{\mathrm{Pooled}}$,使得 $L(\lambda_{\mathrm{Pooled}} \mid \boldsymbol{X}_{\mathrm{Pooled}})$ 最大,即:

$$\hat{\lambda}_{\mathrm{Pooled}} = \arg\max_{\lambda_{\mathrm{Pooled}}} L(\lambda_{\mathrm{Pooled}} \mid \boldsymbol{X}_{\mathrm{Pooled}}) = \arg\max_{\lambda_{\mathrm{Pooled}}} P(\boldsymbol{X}_{\mathrm{Pooled}} \mid \lambda_{\mathrm{Pooled}}) \tag{8}$$

实验结果证明这种模型更具有语种变化的鲁棒性。

3　实验及讨论

实验中,GMM 的高斯混合的数目 M 经验地设置为 64,说话人确认过程采用 cohort 规范化方法[4],这样可以达到打破分类说话人集和冒充说话人之间的平衡以避免夸大说话人确认系统的性能[5]295的目的。

3.1　普通话和四川方言语音库

为了支持说话人确认实验设计了普通话——四川方言语音数据库。该语音库由 22 个说话人(男性:11 人,女性:11 人)的 1 320 条发音(每个说话人 60 条发音)构成。包括普通话发音和四川方言发音各 660 条(每个说话人 30 条普通话发音,30 条四川方言发音)。每个说话人的发音分别包括 20 个句子(普通话和四川方言各 10 个),20 个词组(普通话和四川方言各 10 个)和 20 个 4 位数字组合(普通话和四川方言各 10 个)。

语音数据是在普通办公室环境下录制,录音设备是 Toshiba C9 笔记本电脑(声卡 AC97 标准)和 CD-890MV 头戴式麦克风,录音软件是 Praat,单声道,采样频率 16 kHz,量化精度 16 bit,说话语速为正常语速。录音时并没有刻意避免空调发出的嗡嗡声、电脑 CPU 风扇声以及旁边人私语声。

3.2　单一语种训练高斯混合模型

开始只用 10 个普通话句子训练练说话人的模型,测试时候分别用语音库的普通话数字和四川方言数字,根据公式(5),分别得到的错误率是 6.45% 和 8.76%,如表 1 所示,*M-EER* 表示普通话测试时的错误率,*D-EER* 表示四川方式测试时的错误率。比较得出普通话和四川方言的不匹配导致的性能下降率为 35.74%。

同理,只用 10 个四川方言句子训练说话人的模型,测试时候分别用语音库的普通话数字和四川方言数字,分别得到的错误率是 9.31% 和 6.62%,如表 1 所示。比较得到普通话和四川方言的不匹配导致的性能下降率为 40.53%。

表 1　普通话和四川方言在训练和测试阶段的对比　%

建模方式	*M-EER*	*D-EER*	性能下降率
普通话建模	6.45	8.76	35.74
四川方言建模	9.31	6.62	40.53

从表 1 可以看出用普通话建立的模型时,四川方言测试时的错误率要比普通话测试时的错误率高,同理四川方言建立模型的时候,普通话测试时的错误率比四川方言测试时的错误率高,这说明了 GMM 不仅包括说话人的特性而且还包括训练数据语种的特性。如果只用普通话训练数据来建立模型,说话人的 GMM 包括的更多的是说话人的特性和普通话的语种特性,所以用四川方言测试时的错误率比普通话测时的错误率高。

3.3　按不同比例混合的联合高斯混合模型

将普通话数据和四川方言数据按照比例系数 α 混合来建立普通话和四川方言联合模型,分别用普通话数字和四川方言数字来测试,得到说话人确认系统的结果如表 2 所示。这种建模方法对说话人确认系统带来的性能下降率如图 1 所示。

表 2　按照比例 α 建立的联合模型和确认的结果　%

α	*M-EER*	*D-EER*	性能下降
0/10	6.45	8.76	35.74
2/8	6.96	7.32	5.13
4/6	8.92	9.17	2.79
5/5	10.69	10.40	2.81
6/4	9.31	8.06	15.43
8/2	9.15	8.46	8.06
10/0	9.31	6.62	40.53

从表 2 可以看出(*M-EER* 表示普通话测试时的错误率,*D-EER* 表示四川方式测试时的错误率),当比例系数 α 为

0/10,2/8,4/6 时(普通话训练数据多),用四川方言测试时的错误率要比普通话测试时的错误率高,当比例系数 α 为 5/5,6/4,8/2,10/0 时(四川方言训练数据多),用普通话测试时的错误率比四川方言测试时的错误率高。从图 1 可以看出按照比例 α 混合训练数据来建立普通话和四川方言联合模型可以很大程度上减少上述提出的语种不匹配导致的系统性能下降率,当 α 为 2/8,4/6,5/5,6/4,8/2 时候,说话人确认系统性能下降率分别为 5.13%,2.79%,2.81%,15.43%,8.06%,明显的要小于 α =0/10 时(只用普通话建立模型)的 35.74% 和 α = 10/0 时(只用四川方言建立模型)的 40.53%。尤其是当 α =4/6 时,系统性能下降率达到最低的 2.79%。

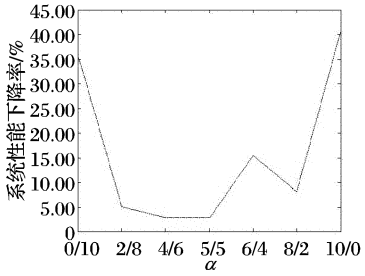

图 1 实验结果

造成这种结果的原因是:GMM 不仅包括说话人的语音特征,而且还包括训练数据的语种特征。而普通话和四川方言在发音、语调和速度方面有着很大的不同,这就决定二者的语种特征有所不同。只用普通话建立模型时,GMM 只包括了说话人的语音特征以及普通话的语种特征,未包括四川方言的语音特征,所以用四川方言测试时的错误率比普通话测试时的错误率高。同理可以推断只用四川方言建模的情形。而按不同时间长度比例混合普通话和四川方言建立模型时,GMM 包括了普通话和四川方言的语种特征,所以分别用普通话和四川方言测试时的错误率相差不大。

实验结果表明,这种按不同时间长度比例混合普通话和四川方言建立联合 GMM 的方法可以更好的解决训练阶段和测试阶段不同语种带来的性能下降,选择用这种方法建立的 GMM 具有更好的鲁棒性和对语种变化的不敏感性。所以这种模型可以更好的用来进行基于普通话和四川方言的说话人确认。

尤其是对本实验中采用的语音库,选择 α =4/6 可以达到很好的效果,但是 α =6/4 时的系统性能下降率为15.43%,相比其他混合比例时的下降率要高很多,原因有待进一步研究。

4 结语

本文提出了按不同时间长度比例 混合普通话和四川方言来建立联合 GMM 的方法,解决了训练阶段和测试阶段的语种的不匹配导致说话人确认系统的性能的急剧下降(从 35.74% 到 40.53%)问题,结果表明这种方法可以有效的解决上述出现的问题。当混合比例 α =4/6 时系统性能下降率达到最低的 2.79%。这说明 GMM 不仅包括说话人的语音特征,而且还包括训练数据的语种特征,在做多语种说话人识别时应该将这个因素考虑在内。以后的研究的重点将把英语以及其他典型中国方言录制进语音库并对多语种带来的说话人确认性能下降率更加进行深入的研究。

参考文献:

[1] 杨澄宇,赵文,杨鉴. 基于高斯混合模型的说话人确认系统[J]. 计算机应用,2001,21(4):7 - 8.

[2] QUATIERI T F. Discrete-time speech signal processing: principles and practice[M]. 北京: 电子工业出版社,2004: 572 - 575.

[3] AUCKENTHALER R, CAREY M J, MASON J S D. Language dependency in text-independent speaker verification[C]// IEEE Acoustics, Speech, and Signal Processing. [S. l.]: IEEE, 2001, 1: 441 - 444.

[4] AUCKENTHALER R, CAREY M, LIOYD-THOMAS H. Score normalisation in a text-independent speaker verification system[J]. Digital Signal Processing, 2000, 10(1): 47 - 48.

[5] MA B, MENG H. English - Chinese bilingual text - independent speaker verification[C]// IEEE Acoustics, Speech, and Signal Processing. [S. l.]: IEEE, 2004, 5: 293 - 295.

[6] REYNOLDS D A. Speaker identification and verification using Gaussian mixture speaker models[J]. Speech Communication, 1995, 17: 97 - 103.

[7] FINAN R A, SAPELUK A T, DAMPER R I. Imposters cohort selection for score normalization in speaker verification[J]. Pattern Recognition letters, 1997, 18: 883 - 887.

(上接第 791 页)

从比较结果可以看出,同样是提取微生物的变差函数纹理特征向量,基于支持向量机的分类效果在回判率和识别率方面明显比 BP 神经网络的分类效果要好。由此可见,基于支持向量机的微生物分类算法对提高分类的准确性有显著的作用。

4 结语

本文讨论了支持向量机在显微图像分类中的应用,利用变差函数提取微生物的纹理特征,基于 RBF 核函数下的 SVM 模式分类原理,建立了微生物显微图像的分类识别模型,该模型基本能正确识别两类微生物。与 BP 神经网络相比,具有比较好的分类精度。这是因为支持向量机不需要设计者的先验知识和经验,并且特别适用于样本数目比较少的分类情况。以上实验结果表明,应用变差函数和支持向量机理论对微生物显微图像进行分类识别这一思路是可行的,该研究为建立自动显微图像识别检测平台奠定了基础,并具有广阔的应用前景。

参考文献:

[1] 王丽亚,李小平. 纹理图像的特征提取和分类[J]. 微电子学与计算机,2005,22(9):96 - 98.

[2] 王积分,阳葵,阎炜. 基于多重分形理论的菌种筛选分类器[D]. 天津: 天津大学,2001.

[3] 杨榕,张荣,孙松. 基于图像处理技术的浮游生物自动分类研究[J]. 计算机仿真,2006,23(5):167 - 170.

[4] 丁乐洪,储炬,庄英萍. 黄青霉形态的显微图像分析研究[J]. 中国抗生素杂志,2003,28(3):131 - 133.

[5] 安金龙,王正欧,马振平. 一种新的支持向量机多分类方法[J]. 信息与控制,2004,33(3):262 - 267.

[6] 李小涛,李纪人,黄诗峰,等. 变差函数和神经网络结合的遥感影像分类方法研究[J]. 国土资源遥感, 2006,67(1):82 - 87.

[7] GU CHUANG, LEE M-C. Semantic video object segmentation and tracking using mathematical morphology and perspective motion model[C]// International Conference on Image Processing (ICIP'97). Washington, DC: IEEE Press, 1997: 514.

[8] CHRISTOPHER J C A. BURGES A. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121 - 167

[9] SEBALD D J, BUCKLEW J A. Support vector machine techniques for nonlinear equalization[J]. IEEE Transactions on Signal Processing, 2000, 48(11): 3217 - 3226.